基于流调度选择的 DCN 动态负载均衡算法 *

李松州，束永安

(安徽大学 计算机科学与技术学院 合肥 230601)

摘 要：针对数据中心网络(data center network，DCN)动态调度导致的负载不均衡问题，提出了基于流调度选择的动态负载均衡(dynamic load balancing based on flow scheduling selection，DLBFSS)算法。该算法首先计算拥塞链路上各条大流的等价最短路径，并删除不满足流带宽需求的路径；然后计算剩余路径的可用吞吐量，选择可用吞吐量最大的路径作为最优调度路径；最后根据大流的带宽和最优路径的负载定义调度的拥塞概率，将拥塞概率作为大流调度选择的依据。实验结果表明，与传统 ECMP（equal-cost multi-path）路由和现有大流调度算法相比，DLBFSS 能够减小网络时延，提高流的带宽利用率，保证了更好的负载均衡。

关键词：数据中心网络；动态负载均衡；软件定义网络；调度选择；拥塞概率

中图分类号：TP393 doi: 10.3969/j.issn.1001-3695.2017.07.0675

Dynamic load balancing algorithm for DCN based on flow scheduling selection

Li Songzhou, Shu Yongan

(*College of Computer Science & Technology Anhui University*, *Hefei* 230601, *China*)

Abstract: In order to solve the problems of poor load balancing caused by dynamic scheduling in data center networks(DCN) , this paper proposed a dynamic load balancing algorithm based on flow scheduling selection(DLBFSS) . Firstly, DLBFSS calculates the equal-cost shortest path for each large flow on the congested link and removes the path with insufficient bandwidth. Then, the available throughput of each remaining path for large flows was calculated, and the path with maximum available throughput was selected as the optimal scheduling path. Finally, the congestion probability of each large flow was defined by considering the flow bandwidth and the load of the optimal path, and the congestion probability was used for large flow scheduling selection. The experimental results show that DLBFSS can reduces the network delay, improves the flow bandwidth utilization and ensures load balancing compared with ECMP(equal-cost multi-path) and typical large flow scheduling algorithms.

Key Words: data center network(DCN); dynamic load balancing; software defined network(SDN); scheduling selection; congestion probability

0 引言

随着计算机网络和数据中心网络（data center network，DCN）的快速发展，需要新的技术来管理和监控网络，软件定义网络（software defined network，SDN）作为一种新的网络架构，通过控制平面与转发平面的解耦，能够实现对网络的集中管理和全局信息的监控[1]。在 DCN 中存在大量时延敏感和高吞吐量的应用[2]，为了解决对 DCN 高带宽和低时延的需求，大多数 DCN 结构选择层次型多根拓扑模型，如 Folded-Clos[3]、Fat-Tree[4]，这些网络模型具有多路径特点，能够为数据流的路由提供多条等价路径，因此在 DCN 中存在大量的带宽资源。

由于受到路由协议的限制，并不能有效的利用当前网络中的带宽资源，路由协议是带宽资源利用效率低下的主要原因之一[5]。因此有效的路由算法的设计和拥塞控制机制，能够实现网络的负载均衡，减小链路的拥塞，提高带宽资源的利用率。数据中心网络属于单一的所有者，这符合 SDN 集中控制的思想[6]，通过集中控制，SDN 控制器能够提供全局的网络视图和实时的网络状态，为路由算法的设计和网络的拥塞控制提供有利条件。本文利用 SDN 网络集中控制的优点提出了 DLBFSS 算法，该算法的主要目标是解决 DCN 中的拥塞控制问题，实现更好的动态负载均衡，提高网络资源的利用率。

1 相关研究

基于 SDN 的数据中心多路径负载均衡算法可以分为动态和静态两种方式。在静态负载均衡中，流在传输期间路径不能发生改变；动态负载均衡允许流在传输期间动态的改变传输路

基金项目：安徽省自然科学基金资助项目（1408085MF125）

作者简介：李松州（1993-），男，安徽阜阳人，硕士研究生，主要研究方向为软件定义网络（SDN）(1659903829@qq.com)；束永安（1966-），男，教授，博士研究生，主要研究方向为无线网络、下一代网络体系结构.

径，能够将拥塞路径上的部分流量调度到其他空闲路径上，实现网络的动态负载均衡。

静态负载均衡算法：等价多路径 ECMP[7]路由算法是一种典型的负载均衡方法，利用网络的多路径特点，通过哈希函数将数据流静态的映射到各条等价路径上，但未考虑网络状态和流的大小，容易导致网络发生拥塞。文献[8]提出的 DLB 算法采用贪婪策略，利用深度优先的方式，每一跳选取空闲带宽最大的链路，由于没有考虑网络的全局负载，不能保证选取路径的质量。文献[9]提出的 CAMOR 是一种拥塞感知的路由算法，在节点之间的多条等价最短路径中，选择可用吞吐量最大且未发生拥塞的路经作为最优路径，但未考虑路径的最小剩余带宽，容易造成链路的过载。

动态负载均衡算法：文献[10]提出了一种单跳的 LABERIO 算法，通过设置一个网络均衡阈值对网络的均衡程度进行判断，并根据所有链路的状态计算网络负载的均衡程度，当均衡程度超过该均衡阈值时，在负载最高的链路上选择一条最大的流进行调度。Wile Sehery 等人[3]提出了 SRL 和 FlowFit 两个负载均衡算法，SRL 作为路由初始化算法，首先随机选择节点对之间的两条等价最短路径，然后将负载最小的一条作为路由；当某条链路出现拥塞时，FlowFit 依次选择该链路上对负载影响最大的流进行调度，直到链路的负载小于拥塞阈值。但 LABERIO 与 FlowFit 这两种动态负载均衡算法都只是选择当前最大的流进行调度，没有考虑其他的大流，选择的不一定是最优对象，容易造成调度路径负载过高或发生拥塞，不利于网络的负载均衡。

针对上述算法中只选择最大流调度的缺点，本文提出了 DLBFSS 算法，根据拥塞链路上各条大流的带宽和调度路径的负载计算大流调度的拥塞概率，利用拥塞概率选择要调度的大流，防止大流的不恰当调度，提高动态负载均衡调度的有效性。

2 基于流调度选择的动态负载均衡方案

本文的动态负载均衡方案由流量测量模块、路由模块和拥塞调度模块组成，将这三个模块实现为 Floodlight[11]控制器的一部分，并依赖于 Floodlight 控制器提供的服务来运行，如图 1 所示。

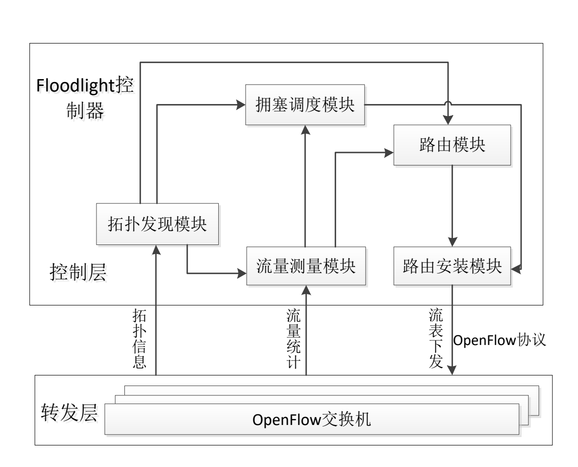


图 1　基于 Floodlight 控制器的动态负载均衡框架

2.1 路由模块

为了比较 FlowFit 与 DLBFSS 两个动态负载均衡算法的拥塞调度效果，本文使用相同的 SRL 算法[3]作为 DLBFSS 的路由模块算法，用于路由的初始化，为每个新到达的流选择路由。过程如下：首先根据网络拓扑信息计算流的所有等价最短路径；然后通过两个不同的哈希函数对流的包头信息进行哈希计算，使用哈希值选择两条等价最短路径；最后利用流量测量模块提供的链路负载信息将负载最小的一条路径作为路由。

2.2 流量测量模块

该模块周期性的向网络中的交换机发送统计请求消息，获得各交换机的流量统计信息，利用这些信息计算网络的负载状态和流量分布，为其他模块提供决策信息。

主要功能：计算各条链路的负载和利用率，统计链路上的大流，检测链路是否发生拥塞，当检测到拥塞链路时，调用拥塞调度模块。

符号表示：本文采用图 $G(V,E)$ 表示网络拓扑，$V=\{v_1,v_2,\dots,v_x\}$ 表示交换机的集合，$E=\{e_1,e_2,\dots,e_y\}$ 示链路的集合，$e_{ij}\in E$ 表示节点 v_i 到节点 v_j 的链路，T 表示周期。

链路负载：链路负载为某一时刻链路的占用带宽，可以通过单位时间内所连端口的收发字节数来测量。将周期 T 内 e_{ij} 所连端口的收发字节数分别为记为 R 和 S，则链路 e_{ij} 的负载定义为：

$$Bw_{ij}=(R+S)/T \tag{1}$$

链路利用率：链路利用率为链路的负载与链路带宽的比值，实验中用 Bc 表示链路带宽，根据已知的链路负载 Bw_{ij}，则链路 e_{ij} 的利用率定义为：

$$Bu_{ij}=Bw_{ij}/Bc \tag{2}$$

大流检测：在 SDN 网络中，OpenFlow 协议对每个流表项都维护了计数器，通过该计数器可以统计流的接收字节数，估计流的带宽。本文用 f_{th} 表示区分大流和小流的阈值，f_s 表示流的带宽，在 t-T 和 t 时刻统计某个流的传输字节数分别为 b_{t-T} 和 b_t，则该流的带宽可以定义为：

$$f_s=(b_t-b_{t-\mathrm{T}})/\mathrm{T} \tag{3}$$

当 $f_s>f_{th}$ 时，认为该流为大流，否则为小流。

链路拥塞检测：在实验中将链路带宽的某一比例(用 β 表示)作为拥塞阈值 B_{th}，记为 $B_{th}=\beta * Bc$，利用拥塞阈值判断链路的拥塞状态。当检测到 $Bw_{ij}>B_{th}$ 或 $Bu_{ij}>\beta$ 时，认为链路 e_{ij} 发生拥塞，调用拥塞调度模块。

2.3 拥塞调度模块

在数据中心网络中存在对带宽要求较高的大流和对时延敏感的小流[12]。当网络发生拥塞时，对小流调度会增加时延开销，而大流对时延要求不高，因此可以选择大流进行调度。

主要功能：选择拥塞链路上各条大流的最优调度路径，计算大流调度的拥塞概率，利用拥塞概率选择要调度的大流。该方法考虑了大流调度时对调度路径的影响，避免造成大流的不恰当调度，有利于网络的负载均衡。

符号表示：如果拥塞链路 e_{ij} 上检测到 K 条大流，记为 $F=\{f_1,f_2\cdots f_k\cdots f_K\}$，第 k 条流 f_k 的等价最短路径有 M 条，第 m 条记为 $p_m^k=\{e_{m1}^k,e_{m2}^k\ldots e_{mi}^k\ldots e_{ml}^k\}$，$e_{mi}^k$ 为路径 p_m^k 的第 i 条链路，Bw_{mi}^k 表示 e_{mi}^k 的链路负载，f_s^k 表示流 f_k 的统计带宽，e_{mi}^k 的可分配带宽记为 $B_{mi}^k=B_{th}-Bw_{mi}^k$。

候选路径的计算：首先使用最短路径算法计算拥塞链路上大流的所有等价最短路径，然后删除不满足流带宽传输需求的路径，即每条路径满足式（4）。

$$f_s^k<\min\{B_{mi}^k \mid e_{mi}^k\in p_m^k\} \tag{4}$$

最终得到流 f_k 的候选路径集合，记为：$P^{k`}=\{p_1^{k`},p_2^{k`}\ldots p_n^{k`}\ldots p_N^{k`}\}$。

最优调度路径的选择：本文选择可用吞吐量最大且满足流带宽传输需求的路径作为最优调度路径。对于流 f_k 的候选路径集合 $P^{k`}$，将候选路径 $p_n^{k`}$ 上每条链路的剩余带宽之和定义为该路径的可用吞吐量，记为 $T(p_n^{k`})$，如式（5）所示。

$$T(p_n^{k`})=\sum_{i=1}^{I}(Bc-Bw_{ni}^{k`}) \tag{5}$$

由公式的定义可知，路径的可用吞吐量越大，拥塞程度就越小，越有利于流的传输。选择 $T(p_n^{k`})$ 值最大的候选路径作为最优调度路径，记为 p_o^k，如式（6）所示。

$$p_o^k=\arg_{p_n^{k`}}\ max\{T(p_n^{k`})\mid n=1,\ldots\ldots,N\} \tag{6}$$

流的调度选择：为使大流的调度更加准确、有效，本文采用最差适应法定义大流调度的拥塞概率，考虑了大流调度对调度路径的影响。将流 f_k 的带宽占最优调度路径瓶颈带宽的比例定义为调度的拥塞概率，记为 C_K，如式（7）所示。

$$C_k=\frac{f_s^k}{\min\{(Bc-Bw_{oi}^k)\mid e_{oi}^k\in p_o^k\}} \tag{7}$$

$\min\{(Bc-Bw_{oi}^k)\mid e_{oi}^k\in p_o^k\}$ 为调度路径 p_o^k 的最小剩余带宽，即 p_o^k 的瓶颈带宽。由公式的定义可知：C_k 的值越小，调度后路径 p_o^k 的拥塞程度就越小，网络负载越均衡。

然后计算出所有大流调度的拥塞概率，利用拥塞概率对大流进行排序，得到大流的调度集合 *Flowlist*，如式（8）所示。

$$Flowlist=\mathrm{sort}\{f_k\mid k=1,\ldots,K\} \tag{8}$$

Flowlist 按 C_k 的值从小到大存放要调度的大流，拥塞概率越小，流被调度的优先级就越高。依次选择拥塞概率最小的大流进行调度，当链路的负载小于拥塞阈值后停止调度。

基于流调度选择的动态负载均衡算法：

```
if( Bw_ij > B_th ) then
  F= e_ij .largeflowlist();
  K= F.length();
  for( f_k  in  F , k=1 to K) do
    p_o^k = f_k .optimalpath();
    C_k = f_k .getprobability ( p_o^k );
    //计算流 f_k 调度的拥塞概率
  end for
  Flowlist = sort{ f_k | k=1,...,K };
  / /按拥塞概率从小到大对大流进行排序存放
  for( f  in  Flowlist ) do
    //依次选择拥塞概率最小的大流进行调度
    reroute( f );
    updatelinkload( f_s );
    //调度后对链路的负载进行更新
    if( Bw_ij < B_th )  break;
    //链路负载小于拥塞阈值后停止调度
  end for
end if
```

由于虚拟化技术和云计算等新兴应用模式的发展，数据中心网络服务器内部通信的东西向流量显著增加[13]，在运行过程中极易造成网络拥塞。当网络发生拥塞时，在缺少流量动态调度的机制中，会使部分链路处于拥挤状态，而有的链路处于空闲状态，导致流量负载不均衡，网络传输时延增加，流的带宽利用率降低，网络整体性能下降。在拥有流量动态调度的机制中，如果未考虑对调度路径的影响，选择了不恰当的流进行调度，也容易产生新的网络拥塞问题，造成流量负载不均衡，不利于网络整体性能的提升。因此本文引入了利用拥塞概率的大流调度选择算法，从上述算法的描述可知，在调度过程中，该方法在减小拥塞链路流量负载的同时，考虑了对调度路径的影响，能够最小化大流调度后调度路径的负载，防止调度路径发生拥塞，在网络负载状态的动态变化过程中，保证了网络中各条链路的流量负载均衡，更有利于流的传输，能够减少网络传输时延，增加流的带宽利用率，从而可以提高网络的整体性能，并通过下文的具体实验对分析进行验证。

3　仿真实验

3.1　仿真场景描述

SDN 环境搭建:实验中采用 Floodlight 控制器作为控制平面，用 Mininet[14]模拟数据平面，Mininet 可以创建包括 Open vSwitch 交换机、链路和终端主机在内的数据平面，实验环境为 64 位 Ubuntu 14.04 操作系统。

数据中心网络模型：使用 Mininet 模拟 Fat-Tree 拓扑作为研究数据中心网络的对象，当 Fat-Tree 拓扑有 k 个 Pod（performance optimization datacenter）时，共有 $5k^2/4$ 台交换机和 $k^3/4$ 台主机。本文取 k=4，创建一个具有 4 个 Pod 的数据中心网络拓扑，则网络中共有 20 台交换机和 16 台主机，其中核心层有 4 台交换机，汇聚层和边缘层分别有 8 台交换机，如图 2 所示。

实验参数：实验中链路的带宽设置为 $1Gbps$，取 $\beta=0.9$，周期 $T=5s$。使用 Iperf 流量生成工具向网络中发送大小不同的流量，为了模拟真实的数据中心网络场景，利用随机流量模型（任意一台主机以相等的概率向网络中的其他主机发送数据流）模拟数据中心网络的流量，取大流和小流区分的阈值为链路容量的 0.5%[3]，即 f_{th} =5Mb/s，根据数据中心网络的流量特征：设

置 90%的流为小流，10%的流为大流。在实验中逐步改变网络的流量负载，测试网络的性能指标，比较负载均衡算法在不同负载下的网络性能。流量负载为主机上链路的平均占用带宽与链路容量的比值，范围为[0,1]，测试 9 组负载，从 0.1 到 0.9。

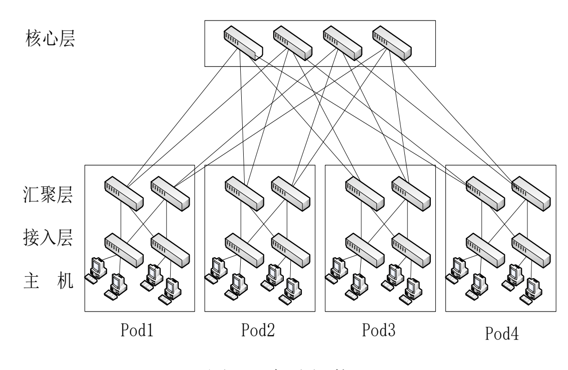


图 2 实验拓扑

3.2 负载均衡评估指标

平均带宽利用率和平均时延是评价网络性能的重要指标，能够反映网络的负载均衡程度。

a)平均带宽利用率。所有流带宽利用率的平均值，单条流的带宽利用率为服务端接收的带宽与客户端发送带宽的比值。

b)平均时延。数据包端到端传输时延的平均值。

3.3 仿真结果与分析

为了验证本文设计的利用拥塞概率进行大流调度选择算法的优越性能，在实验中将 DLBFSS 算法分别与传统的 ECMP 算法和现有的 FlowFit 大流调度算法进行对比，并从流的带宽利用率和网络时延两个方面比较了三种算法的网络性能。其中，ECMP 是目前大多数企业和数据中心网络部署的一种等价多路径负载分担路由算法，采用基于流的哈希方式，利用哈希函数将数据流静态的映射到各条等价路径上，以此来达到流量负载均衡的目的；文献[3]中的 FlowFit 是一种新的典型的数据中心网络流量动态调度算法，通过将拥塞链路上的部分最大流调度到其他负载较小的路径上，减小网络的拥塞程度，实现流在传输过程中的动态负载均。本文的 DLBFSS 对 FlowFit 选择最大流调度的缺点进行了改进，考虑了调度后对网络整体性能的影响，因此选择以上两种类型的负载均衡算法与 DLBFSS 进行对比，并分别在 Floodlight 控制器中实现，仿真结果如图 3、4 所示。

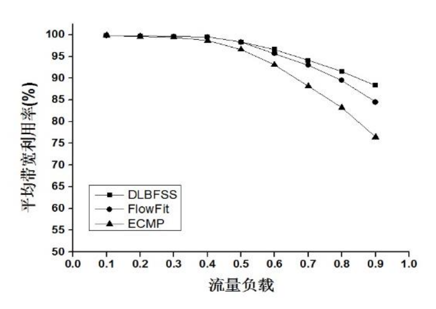


图 3 平均带宽利用率

图 3 显示了三种算法在不同负载下的平均带宽利用率，随着流量负载的增加，DLBFSS 的平均带宽利用率明显高于 ECMP 和 FlowFit，ECMP 的平均带宽利用率最低，当流量负载达到 0.9 时，DLBFSS 的平均带宽利用率比 FlowFit、ECMP 分别提高了 3.8%和 11.9%。ECMP 在路由时未考虑网络的负载，容易使链路发生拥塞，并且缺少流的动态调度机制，导致数据包丢失较多，平均带宽利用率最低。FlowFit 与 DLBFSS 在路由时使用了 SRL 作为路由初始化算法，考虑了网络负载，拥有流的动态调度机制，减少了数据包的丢失，平均带宽利用率高于 ECMP。当检测到链路拥塞时，FlowFit 总是选择当前最大的流进行调度，没有考虑其他的大流，可能造成调度路径的拥塞，DLBFSS 考虑了所有的大流，定义了大流调度的拥塞概率，选择拥塞概率最小的大流进行调度，减小了调度路径拥塞的可能，有利于网络的负载均衡，数据包丢失较少，平均带宽利用率高于 FlowFit。

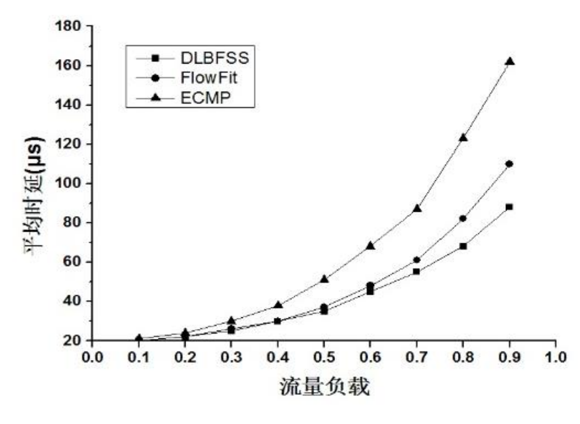


图 4 平均时延

图 4 描述了不同流量负载下平均时延的变化，随着流量负载的增加，三种算法的平均时延不断上升，在相同负载下，ECMP 的平均时延最大，DLBFSS 的平均时延最小，当流量负载达到 0.9 时，DLBFSS 的平均时延比 FlowFit 降低了 22μs、比 ECMP 降低了 74μs。ECMP 通过哈希的方式静态的分配流量，容易将流路由到负载较高的路径上，当链路发生拥塞时，不能对流进行动态的调度，导致数据包的传输时间较长，平均时延最大。FlowFit 与 DLBFSS 在路由初始化时考虑了网络负载，并且能够改变拥塞链路上流的路径，将流调度到其他空闲路径上，减小了网络拥塞，数据包传输时间减少，平均时延低于 ECMP。DLBFSS 在进行拥塞调度时考虑了所有大流的带宽和链路负载，对要调度的各条大流进行了选择，避免了 FlowFit 只选择最大流调度的缺点，网络拥塞程度减小，平均时延低于 FlowFit。

4 结束语

本文提出了 DLBFSS 算法，针对现有的负载均衡动态调度算法只选择最大流调度的缺点进行了改进，优化了大流调度的选择方法。当检测到网络拥塞时，DLBFSS 算法综合考虑了拥塞链路上大流的带宽和等价路径的负载,计算出各条大流的最

优调度路径和拥塞概率，依次选择拥塞概率最小的大流进行调度，直到链路的负载小于拥塞阈值，能够最小化大流调度后调度路径的负载，避免产生新的网路拥塞问题。仿真结果表明，DLBFSS 算法与现有的负载均衡算法 ECMP、FlowFit 相比网络性能更好，解决了 FlowFit 只选择最大流调度的缺点，提高了大流调度的准确性和有效性，实现了更好的动态负载均衡。

参考文献：

[1] Veena S, Rustagi R P, Murthy K N B. Network management and performance monitoring using Software Defined Networks [C]// Proc of the 20th Annual International Conference on Advanced Computing and Communications. 2014: 29-31.

[2] Gholami M, Akbari B. Congestion control in Software Defined Data Center Network through Flow Rerouting [C]// Proc of the 23rd Iranian Conference on Electrical Engineering. 2015: 654-657.

[3] Sehery W, Clancy T C. Load balancing in data center networks with folded-clos architectures [C]// Proc of the 1st IEEE conference on network softwarization. 2015: 1-6.

[4] Jo E, Pan Deng, Liu J, et al. A simulation and emulation study of SDN-based multipath routing for fat-tree data center networks [C]// Proc of Winter Simulation Conference. 2014: 3072-3083.

[5] Wojcik R, Domzal J, Dulinski Z. Flow-Aware Multi-Topology Adaptive Routing [J]. IEEE Communications Letters, 2014, 18 (9): 1539-1542.

[6] 杨洋，杨家海，秦董洪. 数据中心网络多路径路由算法 [J]. 清华大学学报：自然科学版, 2016, 56 (3): 262-268.

[7] Chiesa M, Kindler G, Schapira M. Traffic engineering with equal-cost-multipath: an algorithmic perspective [J]. IEEE//ACM Trans on Networking, 2017, 25 (2): 779-792.

[8] Li Yu, Pan Deng. OpenFlow based load balancing for Fat-Tree networks with multipath support [C]// Proc of the 12th IEEE International Conference on Communications. 2013: 1-5.

[9] Shah S A R, Seok W, Kim J, et al. CAMOR: congestion aware multipath optimal routing solution by using software-defined networking [C]// Proc of International Conference on Platform Technology and Service. 2017: 1-6.

[10] Long Hui, Shen Yao, Guo Minyi, et al. Dynamic load-balanced Routing in OpenFlow-enabled Networks [C]// Proc of the 27th International Conference on Advanced Information Networking and Applications. 2013: 290-297.

[11] Project Floodlight. Floodlight [EB/OL]. [2017-06-05]. http://www. projectfloodlight. org/.

[12] Benson T, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild [C]// Proc of the 10th ACM SIGCOMM Conference on Internet Measurement. 2010: 267-280.

[13] Deng Gang, Gong Zhenghu, Wang Hong. Characteristics research on modern data center network [J]. Journal of Computer Research and Development, 2014, 51 (2): 395-407.

[14] Mininet Team. Mininet [EB/OL]. [2017-06-05]. http://mininet. org/.